IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

Applicant(s): C. Aggarwal et al.
Docket No.: YOR920000430US1
Serial No.: 09/703,174
Filing Date: October 31, 2000
Group: 2176
Examiner: Nathan Hillery

Title: Methods and Apparatus for Intelligent Crawling on the World Wide Web

---

APPEAL BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Applicants (hereinafter "Appellants") hereby appeal the final rejection dated November 8, 2007 of claims 1-27 of the above-identified application.

REAL PARTY IN INTEREST

The present application is assigned to International Business Machines Corp., as evidenced by an assignment recorded October 31, 2000 in the U.S. Patent and Trademark Office at Reel 11290, Frame 0654. The assignee, International Business Machines Corp., is the real party in interest.

RELATED APPEALS AND INTERFERENCES

There are no known related appeals or interferences.

STATUS OF CLAIMS

The present application was filed on October 31, 2000 with claims 1-27. Claims 1-27 are currently pending in the application. Claims 1, 10 and 19 are the independent claims.

Each of claims 1-27 stands finally rejected under 35 U.S.C. §103(a). Claims 1-27 are appealed.

STATUS OF AMENDMENTS

There have been no amendments filed subsequent to the final rejection.

SUMMARY OF CLAIMED SUBJECT MATTER

Independent claim 1 is directed to a computer-based method of performing document retrieval in accordance with an information network. The method includes a step of initially retrieving one or more documents from the information network that satisfy a user-defined predicate. The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one. The method also includes steps of collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed and using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network. The statistical information using step further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

An illustrative embodiment of the claimed computer-based method of performing document retrieval in accordance with an information network is shown in FIG. 2 and described in the specification at, for example, page 8, lines 10-23. The method includes a step of initially retrieving one or more documents from the information network that satisfy a user-defined

predicate (see the specification at, for example, page 8, line 23, to page 9, line 10). The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one (see the specification at, for example, page 4, line 22, to page 5, line 5). The method also includes steps of collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed (see the specification at, for example, page 9, lines 11-15, with respect to steps 260 and 290 in FIG. 2) and using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network (see the specification at, for example, page 9, lines 15-18, with respect to steps 300 and 310 in FIG. 2). The statistical information using step further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user (see the specification at, for example, page 5, lines 6-9, and page 6, lines 15-18).

Independent claim 10 is directed to an apparatus for performing document retrieval in accordance with an information network; the apparatus comprises at least one processor. The at least one processor is operative to initially retrieve one or more documents from the information network that satisfy a user-defined predicate. The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one. The processor is further operative to collect at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and use the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network. The statistical information using operation further comprises

learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

In an illustrative embodiment shown in FIG. 3, a claimed apparatus (computer system 10) for performing document retrieval in accordance with an information network (world wide web 20) comprises at least one processor (CPU 12). The at least one processor is operative to initially retrieve one or more documents from the information network that satisfy a user-defined predicate (see the specification at, for example, page 8, line 23, to page 9, line 10). The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one (see the specification at, for example, page 4, line 22, to page 5, line 5). The processor is further operative to collect at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed (see the specification at, for example, page 9, lines 11-15, with respect to steps 260 and 290 in FIG. 2) and use the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network (see the specification at, for example, page 9, lines 15-18, with respect to steps 300 and 310 in FIG. 2). The statistical information using operation further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user (see the specification at, for example, page 5, lines 6-9, and page 6, lines 15-18)..

Independent claim 19 is directed to an article of manufacture for performing document retrieval in accordance with an information network, comprising a machine readable medium containing one or more programs. The one or more programs when executed implement a step of initially retrieving one or more documents from the information network that satisfy a user-defined predicate. The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves

the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one. The one or more programs when executed also implement steps of collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed and using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network. The statistical information using step further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

An illustrative embodiment of an article of manufacture for performing document retrieval in accordance with an information network, comprising a machine readable medium containing one or more programs, is described in the specification at, for example, page 7, line 25, to page 8, line 9. The one or more programs when executed implement a step of initially retrieving one or more documents from the information network that satisfy a user-defined predicate (see the specification at, for example, page 8, line 23, to page 9, line 10). The initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one (see the specification at, for example, page 4, line 22, to page 5, line 5). The one or more programs when executed also implement steps of collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed (see the specification at, for example, page 9, lines 11-15, with respect to steps 260 and 290 in FIG. 2) and using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network (see the specification at, for example, page 9, lines 15-18, with respect to steps 300 and 310 in FIG. 2). The statistical information using step further comprises learning a linkage structure from at least a portion of

the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user (see the specification at, for example, page 5, lines 6-9, and page 6, lines 15-18).

GROUND OF REJECTION TO BE REVIEWED ON APPEAL

1. Claims 1-8, 10-17 and 19-26 are rejected under 35 U.S.C. §103(a) as being unpatentable over S. Chakrabarti et al., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," Computer Networks, 25 pages, 1999 (hereinafter "Chakrabarti") in view of U.S. Patent No. 6,529,901 (hereinafter "Chaudhuri").

2. Claims 9, 18 and 27 are rejected under 35 U.S.C. §103(a) as being unpatentable over Chakrabarti in view of S. Chakrabarti et al., "Distributed Hypertext Resource Discovery Through Examples," Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, pp. 375-386, 1999 (hereinafter "Ch2").

ARGUMENT

1. Rejection of claims 1-8, 10-17 and 19-26 under §103(a) over Chakrabarti and Chaudhuri.

Appellants respectfully submit that the combination of Charkrabari and Chaudhuri fails to teach or suggest all the claim limitations that the final Office Action asserts it does, and that there is no cogent motivation for combining the reference teachings, as asserted by the final Office Action, to reach the claimed invention.

Claim 1 includes a limitation wherein a computer-based method of performing document retrieval in accordance with an information network includes a step of collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed. Aggregate statistical information and predicate-specific statistical information are respectively described in the present specification as information maintained for all retrieved documents and information maintained for the subset of the retrieved documents which satisfy a given predicate. See, for example, page 8, lines 18-20, and page 10, lines 15-25.

In formulating the rejection of claim 1 on page 4 of the final Office Action, the Examiner concedes that Chakrabarti fails to disclose the aforementioned limitation of claim 1 wherein at least a set of aggregate statistical information and a set of predicate-specific statistical information are collected about the one or more retrieved documents as the one or more retrieved documents are analyzed. The Examiner instead argues that this limitation is met by column 19, lines 35-63 of Chaudhuri.

Appellants have reviewed the relied-upon portion of Chaudhuri and have found no teaching or suggestion directed to collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed. Rather, the relied-upon portion of Chaudhuri is directed toward a technique, MNSA, for "determining if the existing set of statistics contains an essential set of statistics." (Chaudhuri, column 19, lines 35-36, with emphasis added).

Moreover, as noted above, claim 1 includes a limitation directed toward collecting a set of aggregate statistical information, e.g., information maintained for all retrieved documents. The Examiner apparently contends that this limitation is met by Chaudhuri's disclosure that "Aggregation (GROUP BY or SELECT DISTINCT) clauses can be handled by associating a selectivity variable that indicates the fraction of rows in the table with distinct values of the column(s) in the clause." Appellants respectfully submit that a disclosure of a manner in which GROUP BY or SELECT DISTINCT clauses may be handled fails to teach or suggest a limitation directed toward collecting a set of information maintained for all retrieved documents.

Likewise, claim 1 includes a limitation directed toward collecting a set of predicate-specific statistical information, e.g., information maintained for the subset of the retrieved documents which satisfy a given predicate. The Examiner apparently contends that this limitation is met by Chaudhuri's disclosure that "MNSA guarantees inclusion of an essential set of the query only as long as the selectivity of predicates in the query is between g and 1-g." Appellants respectfully submit that a disclosure that an algorithm guarantees inclusion of an essential set of a query only so long as the selectivity of predicates in the query is within a certain range fails to teach or suggest a limitation directed toward collecting a set of information maintained for the subset of the retrieved documents which satisfy a given predicate.

Accordingly, Appellants respectfully submit that the relied-upon portions of Chaudhuri

fail to remedy the fundamental deficiencies of Chakrabarti with regard to claim 1. Accordingly, Appellants respectfully submit that the combined teachings of Chakrabarti and Chaudhuri fail to render the limitations of claim 1 obvious, as alleged by the Examiner.

Moreover, even if it were possible to combine Chakrabarti and Chaudhuri so as to reach the limitations of claim 1, the Examiner has failed to proffer a legally sufficient explanation for why one having skill in the art would have done so. The Examiner asserts in the final Office Action at page 4, last paragraph, that "[b]ecause both Chakrabarti et al. and Chaudhuri et al. teach methods of collecting statistics, it would have been obvious to one skilled in the art to substitute one method for the other to achieve the predictable result of collecting aggregate and predicate-specific statistics." Examiner's explanation is a conclusory statement of the sort rejected by both the Federal Circuit and the U.S. Supreme Court. See KSR v. Teleflex, 127 S.Ct. 1727, 1741, 82 USPQ2d 1385, 1396 (U.S., Apr. 30, 2007), quoting In re Kahn, 441 F. 3d 977, 988 (Fed. Cir. 2006) ("[R]ejections on obviousness grounds cannot be sustained by mere conclusory statements; instead, there must be some articulated reasoning with some rational underpinning to support the legal conclusion of obviousness.")

More specifically, the statement above is using the benefit obtained from a combination as a motivation for that combination; this is impermissible hindsight. In order to avoid the improper use of a hindsight-based obviousness analysis, particular findings must be made as to why one skilled in the relevant art, having no knowledge of the claimed invention, would have combined the teachings of Chakrabarti and Chaudhuri in the claimed manner. See, e.g., In re Kotzab, 217 F.3d 1365, 1371, 55 USPQ2d 1313, 1317 (Fed. Cir. 2000). The Examiner's conclusory statements do not adequately address the issue of motivation to combine references. "It is improper, in determining whether a person of ordinary skill would have been led to this combination of references, simply to '[use] that which the inventor taught against its teacher.'" In re Sang-Su Lee, 277 F.3d 1338, 1344 (Fed. Cir. 2002) (quoting W.L. Gore v. Garlock, Inc., 721 F.2d 1540, 1553, 220 USPQ 303, 312-13 (Fed. Cir. 1983)).

Independent claims 10 and 19 contain limitations similar to those of claim 1 and are believed patentable for at least the reasons identified above with reference to claim 1.

Dependent claims 2-8, 11-17 and 20-26 are believed patentable for at least the reasons set forth above with reference to the independent claim from which each depends. Furthermore, one

or more of these claims defines independently patentable subject matter.

2. Rejection of claims 9, 18 and 27 under §103(a) over Chakrabarti and Ch2.

As a preliminary matter, Appellants respectfully note that the final Office Action indicates, in the third paragraph of page 7, that claims 9, 18 and 27 are rejected under 35 U.S.C. 103(a) as being unpatentable over Chakrabarti as applied to claims 1-8, 10-17, and 19-26 above, and further in view of Ch2. However, as noted previously, claims 1-8, 10-17 and 19-26 were rejected as being unpatentable over Chakrabarti in view of Chaudhuri. Moreover, the Examiner has failed to indicate the manner in which Ch2 is believed to overcome the acknowledged failure of Chakrabarti to disclose the limitations of the independent claims. Appellants therefore believe that the rejections of claims 9, 18 and 27 should be characterized as being over the combination of Chakrabarti, Chaudhuri and Ch2, rather than over Chakrabarti and Ch2.

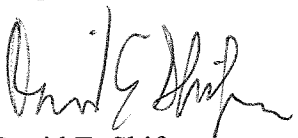
Moreover, in addition to being patentable for at least the reasons set forth above with reference to the independent claim from which each depends, dependent claims 9, 18 and 27 define independently patentable subject matter. Specifically, dependent claims 9, 18 and 27 each contain a limitation directed to statistical information collection using one or more uniform resource locator tokens in the one or more retrieved web pages. In illustrative embodiments described in the present specification at, for example, page 6, lines 8-12, and page 11, lines 19-23, feature extraction mechanisms from the tokens inside the candidate URL may be used. In general, URL names of web pages contain highly relevant information, since web page and server names are usually not chosen randomly. For example, the word "ski" in the URL is highly suggestive that web pages may be related to skiing. Thus, a priority calculation operation may comprise determining a set of URL tokens which have a higher representation in predicate-specific statistics than in aggregate statistics is determined then finding a percentage of tokens in said set which are contained in the tokenized representation of the URL string for a particular candidate URL.

The Examiner contends that Chakabarti fails to teach or even suggest the limitations of claims 9, 18 and 27. Instead, the Examiner relies on Ch2 at page 382, column 1, lines 29-37, which teach that a crawler typically "scans each fetched page for outgoing hyperlink URL's. However, other strategies are also known. E.g., if the URL is of the form http://host/path, the

crawler may truncate components of path and try to fetch these URLs." Appellants respectfully contend that the teachings of Ch2 regarding analyzing URL components in order to determine other pages to be crawled fails to teach or suggest the limitations of claims 9, 18 and 27 directed to directed to using one or more uniform resource locator tokens in the one or more retrieved web pages in statistical information collection.

In view of the above, Appellants believe that claims 1-27 are in condition for allowance, and respectfully request the withdrawal of the §103(a) rejections.

Respectfully submitted,



Date: May 12, 2008

David E. Shifren
Attorney for Applicant(s)
Reg. No. 59,329
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560
(516) 759-2641

CLAIMS APPENDIX

1. A computer-based method of performing document retrieval in accordance with an information network, the method comprising the steps of:

initially retrieving one or more documents from the information network that satisfy a user-defined predicate, wherein the initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one;

collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and

using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network, wherein the statistical information using step further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

2. The method of claim 1, wherein the user-defined predicate specifies content associated with a document.

3. The method of claim 1, wherein the statistical information collection step uses content of the one or more retrieved documents.

4. The method of claim 1, wherein the statistical information collection step considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

5. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

6. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

7. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

8. The method of claim 1, wherein the information network is the world wide web and a document is a web page.

9. The method of claim 8, wherein the statistical information collection step uses one or more uniform resource locator tokens in the one or more retrieved web pages.

10. Apparatus for performing document retrieval in accordance with an information network, the apparatus comprising:

at least one processor operative to: (i) initially retrieve one or more documents from the information network that satisfy a user-defined predicate, wherein the initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one; (ii) collect at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and (iii) use the collected statistical information to automatically determine further document

retrieval operations to be performed in accordance with the information network, wherein the statistical information using operation further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

11. The apparatus of claim 10, wherein the user-defined predicate specifies content associated with a document.

12. The apparatus of claim 10, wherein the statistical information collection operation uses content of the one or more retrieved documents.

13. The apparatus of claim 10, wherein the statistical information collection operation considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

14. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

15. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

16. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

17. The apparatus of claim 10, wherein the information network is the world wide web

and a document is a web page.

18. The apparatus of claim 17, wherein the statistical information collection operation uses one or more uniform resource locator tokens in the one or more retrieved web pages.

19. An article of manufacture for performing document retrieval in accordance with an information network, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

initially retrieving one or more documents from the information network that satisfy a user-defined predicate, wherein the initial document retrieval operation is performed without assuming a specific model of a linkage structure such that the initial document retrieval operation retrieves the one or more documents without assuming that a relationship exists between a feature of a first one of the one or more documents and a feature of at least another one of the one or more documents that links to the first one;

collecting at least a set of aggregate statistical information and a set of predicate-specific statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and

using the collected statistical information to automatically determine further document retrieval operations to be performed in accordance with the information network, wherein the statistical information using step further comprises learning a linkage structure from at least a portion of the collected statistical information with each successive document retrieval operation such that the learned linkage structure is available for use in performing subsequent document retrieval operations requested by a user.

20. The article of claim 19, wherein the user-defined predicate specifies content associated with a document.

21. The article of claim 19, wherein the statistical information collection step uses content of the one or more retrieved documents.

22. The article of claim 19, wherein the statistical information collection step considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

23. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

24. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

25. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

26. The article of claim 19, wherein the information network is the world wide web and a document is a web page.

27. The article of claim 26, wherein the statistical information collection step uses one or more uniform resource locator tokens in the one or more retrieved web pages.

EVIDENCE APPENDIX

None

RELATED PROCEEDINGS APPENDIX

None